IN THE NAME OF GOD

AN INTELLIGENT AGENT-ORIENTED STRUCTURE FOR TEXT INFORMATION RETRIEVAL

BY
MAZIAR SALEHI

THESIS

SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN
PARTIAL FULFILLMANT OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (M.Sc)

IN
COMPUTER ENGINEERING-
ARTIFICIAL INTELLIGENCE AND ROBOTICS
SHIRAZ UNIVERSITY
SHIRAZ, IRAN

EVALUATED AND APPROVED BY THE THESIS COMMITTEE AS: **EXCELLENT**

A. Towhidi .....**A. TOWHIDI, Ph.D., ASSISTANT PROF. OF COMPUTER ENGINEERING (CHAIRMAN)**

M. Zolghadri .....**M. ZOLGHADRI JAHROMI, Ph.D., ASSISTANT PROF. OF COMPUTER ENGINEERING (CHAIRMAN)**

.....**M. H. SADREDDINI, Ph.D., ASSISTANT PROF. OF COMPUTER ENGINEERING**

November 2000

۴۶.۵۴

In memory of my father,
The mount of patience and endurance;
And to my mother,
The foaming spring of affection and love;
And to all those whom I care for,
Without knowing it.

Great discoveries and achievements invariably involve the cooperation of many minds.

Alexander Graham Bell

I hear and I forget. I see and I believe. I do and I understand.

Confucius

٢٩.٥٢

ACKNOWLEDGMENT

The author wishes to express the deepest and sincere thanks and appreciation to his supervisors *Dr. A. Towhidi* and *Dr. M. Zolghadri Jahromi* for their supervision, guidance, contributions and encouragement and for kindly and patiently giving many hours of their time.

The author would also like to express his gratitude to the member of the thesis committee, *Dr. M. H. Sadreddini* for his valuable suggestion and criticism. Thanks are also due to *Dr. Khodaparasti*, from department of foreign languages, for his guidance and cooperation.

Thanks are also due to all of my entire colleague students and friends whose constant encouragement have improved the quality of this work and have modestly requested not to be mentioned, but especially *Mr. A. Sajjadieh* deserves to be mentioned here.

Finally, the author expresses his warmest and most passionate regards to his family who are the dearest moral supporters.

ABSTRACT

AN INTELLIGENT AGENT-ORIENTED STRUCTURE FOR TEXT INFORMATION RETRIEVAL

BY
MAZIAR SALEHI

As the volume of information available on the Internet and corporate intranets is increasing, there is a growing need for tools to help people better find, filter and manage these resources. Artificial intelligence methods, by their adaptable and flexible properties, can help us handle the complexity of these new tools.

This thesis deals with modeling information retrieval (IR) systems using an agent-oriented (AO) approach. The AO approach, which is a new paradigm in software engineering, tries to model software modules as autonomous, intelligent creatures, which cooperate with each other to achieve the desired goal(s). In this work, based on modeling techniques proposed by other researchers, a new agent-oriented technique is proposed, and a prototype system is designed using this method.

In the field of text retrieval, different retrieval methods are investigated and their performances are evaluated using standard test collections (CRANFIELD, MEDLINE). Latent Semantic

Indexing technique is used in this work to extract the associative relationship between terms and documents. A genetic algorithm is also developed to enhance the performance of the matching between queries and documents.

TABLE OF CONTENTS

CONTENT PAGE

CONTENT PAGE

CONTENT PAGE

CONTENT PAGE

LIST OF TABLES

TABLE PAGE

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

FIGURE PAGE

FIGURE PAGE

CHAPTER I

INTRODUCTION

The growth of the Internet and the availability of enormous volumes of data (information explosion) have necessitated intense interest in techniques to assist the user in locating data of interest. Internet has covered over 350 million pages of data till 1998 [1] and is predicted to reach over several billions in first decade of 21st century. The digital library effort is also progressing with the goal of migrating from the traditional book environment to a digital library environment.

Information retrieval proposes methods for organizing and modeling information and user's needs. IR methods basically are funded on information theory and they try to automatically perform all of the retrieval processes. Initial works are performed on text items (news, reports...) but in later years other information items (image, audio, etc) are also involved. However, most of information still consists of huge amount of text items, so text retrieval approaches are still as important portion of IR.

Several models are proposed for intelligent IR with the aid of statistics, probability theory and machine learning techniques. Statistical model such as vector space, which empowered by information theory, is one of the initial successful models.